Human Action Recognition Using Pyramid Histograms of Oriented Gradients and Collaborative Multi-task Learning

Zan Gao[1,2], Hua Zhang[1,2], An-An Liu[3], Yan-bing Xue[1,2] and Guang-ping Xu[1,2]
[1] Key Laboratory of Computer Vision and System, Ministry of Education, Tianjin University of Technology, Tianjin, 300384, P.R. China
[2]Tianjin Key Laboratory of Intelligence Computing and Novel Software Technology, Tianjin University of Technology, Tianjin, 300384, P.R. China
[3]School of Electronic Information Engineering, Tianjin University, Tianjin, 300172, P.R. China
[e-mail: anan0422@gmail.com]
*Corresponding author: AnAn Liu

Received November 6, 2013; revised December 31, 2013; accepted January 9, 2014; published February 28, 2014

Abstract

In this paper, human action recognition using pyramid histograms of oriented gradients and collaborative multi-task learning is proposed. First, we accumulate global activities and construct motion history image (MHI) for both RGB and depth channels respectively to encode the dynamics of one action in different modalities, and then different action descriptors are extracted from depth and RGB MHI to represent global textual and structural characteristics of these actions. Specially, average value in hierarchical block, GIST and pyramid histograms of oriented gradients descriptors are employed to represent human motion. To demonstrate the superiority of the proposed method, we evaluate them by KNN, SVM with linear and RBF kernels, SRC and CRC models on DHA dataset, the well-known dataset for human action recognition. Large scale experimental results show our descriptors are robust, stable and efficient, and outperform the state-of-the-art methods. In addition, we investigate the performance of our descriptors further by combining these descriptors on DHA dataset, and observe that the performances of combined descriptors are much better than just using only sole descriptor. With multimodal features, we also propose a collaborative multi-task learning method for model learning and inference based on transfer learning theory. The main contributions lie in four aspects: 1) the proposed encoding the scheme can filter the stationary part of human body and reduce noise interference; 2) different kind of features and models are assessed, and the neighbor gradients information and pyramid layers are very helpful for representing these actions; 3) The proposed model can fuse the features from different modalities regardless of the sensor types, the ranges of the value, and the dimensions of different features; 4) The latent common knowledge among different modalities can be discovered by transfer learning to boost the performance.

Keywords: Action Recognition, collaborative multi-task learning, PHOG, depth

This research was supported by the National Science Foundation of China, under grant No.61202168, 61201234, 61100124, 21106095.
http://dx.doi.org/10.3837/tiis.2014.02.009

1. Introduction

Recently, human action recognition has been become research hotspot in computer vision and machine learning domain, and widely applied in many areas, such as surveillance video analysis, man-machine interaction and video semantic retrieval etc. In the past decades, a lot of action recognition algorithms [1-6] have been proposed. In these approaches, motion history image [1-2] and spatio-temporal interest points methods [3-6] have been extensively used for representing and recognizing actions. In order to obtain accurate MHI, the target need to be segmented by various detection methods, or the background pixels will affect the construction of MHI. Similar, when we extract spatio-temporal interest points, some interest points will locate in the background, which are noise for target interest point pixels, thus, researchers still want to obtain the target contour to filter noise interest points. However, in traditional videos, it is nontrivial to quickly and reliably detect, segment and track human body, especially in the dark night where the visual contents are unrecognizable or even invisible, and most of the-state-of-the-art approaches will be failed. Specially, after obtaining MHI, different descriptors are borrowed to represent human motions, such as 7-Hu-Moment [7], Gabor [8], but what kinds of descriptors are suitable and robust for representing human motions? In addition, these the-state-of-art methods are assessed on well-known benchmarks, such as Weizmann [2] and KTH [4] datasets, and they can obtain satisfying recognition accuracy, but these datasets just include RGB information, and what the performance will be when these descriptors are applied into depth channel.

With the development of imaging technology, such as Microsoft Kinect and Leap Motion sensors, it has become possible to capture both color image and depth sequences simultaneously with cheap device. and **Fig. 1** shows the outputs of Kinect sensor in which RGB image and depth information are given. From it, we can observe that the depth sequence can supply much more information, such as additional human body shape and motion information, at the same time, the background pixels can be filtered easily by distance, and the target could be segmented and located well and truly. Thus, the depth will be very helpful for our task, and is a powerful supplement for RGB channel. In fact, some researchers [9-10] have tried to adopt the depth information to human action recognition. For example, space-time volume in depth and simple descriptors [9] were constructed and extracted to represent actions, and then approximate string matching (ASM) was used to recognize the action in depth; A bag-of-3D-points or 3D silhouettes method to represent postures by sampling 3D points from depth maps was proposed in Li et al. [10], and then an action graph was borrowed to model these points to perform action recognition. Although these descriptors obtain good performance, they focus on depth channel captured by Kinect sensor, and ignore the RGB information.

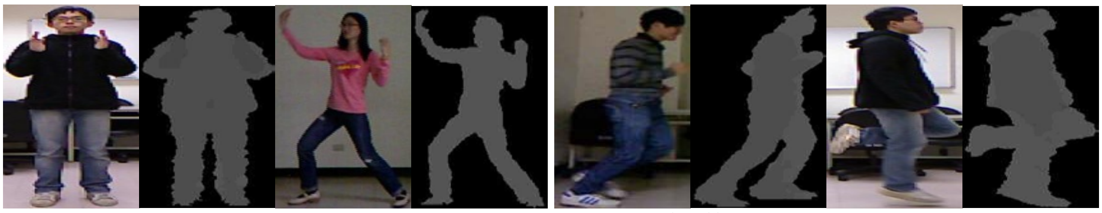


Fig. 1. RGB Images and Depth Maps of different actions

Thus, in this paper, we will first design some descriptors which will suitable for both RGB and depth modalities, and then evaluate these descriptors with different kind of classification models on both RGB and depth modalities, at the same time, we will compare them with classical descriptors which had been utilized in KTH and Weizmann datasets. In addition, with multi-modality features, we also propose a collaborative multi-task learning method for model learning and inference based on transfer learning theory. For this purpose, we first design two kinds of motion history maps for depth and RGB channels: 1) Depth motion history image (**DMHI**) is generated by searching the maximum and minimum motion energy of each pixel in consecutive frame, and then their differences are identified as the **DMHI**; 2) RGB and depth motion history image (**RDMHI**), which is filtered and limited by depth information, is produced to characterize corresponding action categories. After that, average value in hierarchical block, gist and pyramid histograms of oriented gradients descriptors for depth and RGB channels are proposed to represent human motion respectively. In order to evaluate and analyze our proposed descriptors, KNN, SVM with linear and RBF kernels, SRC and CRC model are utilized, at the same time, a public and challenge DHA action dataset, which includes both depth and RGB information together, are employed. **Fig.2** displays the general framework of our approach.

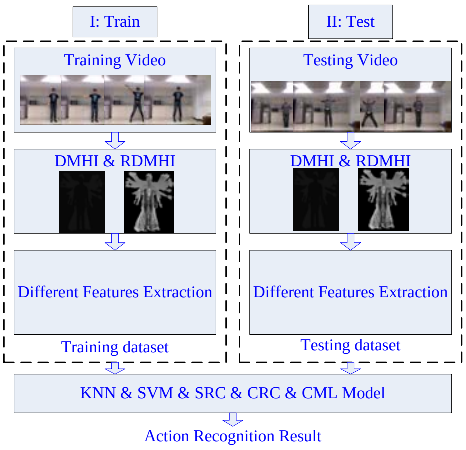

I: Train
II: Test
Training Video
Testing Video
DMHI & RDMHI
DMHI & RDMHI
Different Features Extraction
Different Features Extraction
Training dataset
Testing dataset
KNN & SVM & SRC & CRC & CML Model
Action Recognition Result

Fig. 2. the overall framework of proposed scheme

Large-scale experimental results disclose that in our proposed DMHI_PHOG and RDMHI_PHOG descriptors, the neighbor gradients information and pyramid layers are very useful for our task, whose accuracies are the best on both depth and RGB channels. In addition, we also investigate the performance of our descriptors further by combining these descriptors on DHA dataset, and observe that the performances of combined descriptors are much better than just sole descriptor. What is more, with multi-modality features, collaborative multi-task learning is very helpful for our task. The large scale comparison experiments on public DHA dataset, show the superiority of the proposed method which outperforms the state-of-the-art methods.

The rest of the paper is structured as follows. In Section II and III, we respectively present the related work and motion history maps, and then feature representation and collaborate multi-task learning are given in detail respectively. The experimental results are detailed in Section VI. The conclusions are given at last.

2. Related Work

Human action recognition is a challenging problem because of the high variability of appearances, potential occlusions and shapes. Recently it has obtained increasing attention owing to its wide range of applications in surveillance video analysis, man-machine interaction and video semantic retrieval etc. In the last few decades, a lot of feature representation methods have been developed for recognizing actions from video sequences based on color/RGB cameras. For example, Sequences of human silhouettes are employed to model both spatial and temporal characteristics of human actions [1]. In [1], motion energy images (MEIs) and motion history images (MHIs) are formed by temporally accumulated silhouettes, and then Seven Hu moments [7] and Gabor [8] features are extracted from both MEIs and MHIs to serve as action descriptors; Gaussian mixture models (GMM) is utilized in [11] to capture the distribution of the moments of silhouette sequences. In addition, motion flow patterns to represent human actions are proposed in [12-14]. In details, optical flows [13] are calculated for the entire image by matching consecutive video frames, and then the motion patterns [12] or the estimated motion parameters [14] are utilized for action representation. However, in the real-world, we project three-dimensional motion onto the two-dimensional image plane, it will be ambiguous.

Recently, in object recognition task, local interest points approaches, which are much robust for posture, illumination, occlusion, deformation and cluttered background than that of global appearance descriptions, are very popular, thus, researcher also designed and developed spatio-temporal interest points for action recognition in video, which achieve the state-of-the-art performances in activity recognition. The success reasons of spatio-temporal local features-based methods are that these features are much more distinctive and descriptive. These methods include Cuboid [3], Harris3D [5], MoSIFT [6] and HOG3D [15]. In [3], Dollar et al. focus on temporal domain and put away the spatial constraints, thus, it can detecte periodic frequency components by employing Gabor filters on the temporal dimension. Laptev et al. [5] proposed a 3D Harris corner detector by extending 2D Harris corner detectors, to detect compact and distinctive interest points, which have high intensity variations in both temporal and spatial dimensions. Chen et al. [6] utilized the famous local interest points algorithm to detect visually remarkable areas in the spatial domain, and then these candidated interest points were reserved with the motion constrain, in which an ‘enough’ amount of optical flow around each distinctive points are employed. Although slightly different from each other, these methods share the common feature extraction and representation framework, which involves detecting local extremes of the image gradients and describing the point using histogram of oriented gradients (HOG) and histogram of optic flows (HOF).

Although many algorithms have been proposed for this task, but the previously related works mainly focus on analyzing video sources captured by RGB camera, and have achieved good performance, but what will the performance be when these descriptors are applied into depth channel? Thus, some researchers paid attention to action recognition on depth dataset, for example, Lin *et al.* [9] constructed the space-time volume on depth channel, and then proposed different kinds of descriptors, after that, approximate string matching was employed as the classifier; Wang *et al.* [16] employed the depth and skeleton point information, and constructed the actionlet ensemble model to recognize the actions; A Bag-of-3D-Points was proposed in [10] to represent postures by sampling 3D points from depth maps, and then they employed action graph to model these points to realize action recognition. Their experimental results on MSR Action3D dataset demonstrated that 3D silhouettes from depth sequence are much helpful for action recognition than 2D silhouettes. Megavannan *et al.* [17] also record a depth action dataset, and then constructed different motion history images, after that, different features are cascaded, and SVM classification is trained and employed; Although these algorithms were estimated on depth action dataset, there are still confusions about how to employ depth information, and what kinds of descriptors are suitable for depth channel or for both RGB and depth channels? Secondly, since RGB and depth images represent one scene in different modalities, they are complementary to each other and it will benefit human action recognition by fusing both for discriminative feature representation and model construction. In fact, in different research domains, the fusion of multi-modalities features or multi-view features have attracted the attentions of many scientists. For example, in web image search [18-20], video semantic annotation or tagging [21-24], 3D Object Retrieval [25-28], target tracking [29] and multi-view object classification [30-34], authors had discussed the importance of fusion of multi-modalities features or multi-view features, and experiments also showed its was very helpful for the tasks in different research domains. Thus, we will first assess the performances when these descriptors in RGB and Depth channels are combined. Further, with the features from multiple modality resources, we also propose a collaborative multi-task learning based on transfer learning for human action recognition to assess the importance of the fusion of multi-modality features.

In addition, for the algorithms evaluation, most above algorithms are just assessed by a kind of classification model, but it is not adequate. For example, after extracting different kind of features, all researchers [3-6,17] adopt SVM models to recognize human action; Approximate string matching [9] and graph model [10] are employed to identify human motion; In Bobick and Davis [1], similarity matching schemes were employed; What is worse, most current methods are highly dependent on dataset and therefore the generalization ability is severely constrained. To solve this problem, some authors have proposed a model-free method for human action recognition via sparse representation. For example, Authors [35-42] extracted different kind of features for each action, and then employed sparse representation based classification algorithm directly without any changing. SRC [41] has been proposed firstly for face recognition, in which a testing sample is reconstructed and represented by all the training samples, after that, impulse function is designed for each class and representation, and then the minimum representation error is adopted to classify the testing sample. Similar to SRC, the philosophy of the proposed method in [35-40] is to decompose each video sample containing one kind of human actions as a ℓ_1 sparse linear combination of several video samples containing multiple kinds of human actions, and it has achieved good performance. The reason of obtaining success is that the point's neighborhood structure is utilized fully, and can supply better similarity measures among the testing data and all the training samples. After that, Zhang *et al.* [41] discussed the role of ℓ_1-norm and ℓ_2-norm respectively, and then concluded that the sparsity in SRC was not so important, and collaborative representation played much more important roles. Thus, what will be happened when these descriptors are assessed by mode-free models and traditional, constrained classification algorithms depended on dataset?

3. Motion History Image for RGB and Depth Modalities

In order to represent human motion, human silhouettes of each frame need to be accumulated and encoded firstly, thus, we construct human motion maps for RGB and depth channel respectively, and the details will be given as follows.

3.1 MHI for RGB Modality

To describe human motion, motion history images (**MHI**) [1], where moving human silhouettes are accumulated and encoded, has been widely employed, and achieved good performance. However, Bobick and Davis [1], firstly detected or segmented targets in RGB video, but it is difficult to come true it in real conditions. Thus, in the construction of **MHI**, we often make no discrimination for all pixels in RGB image, by this way, a lot of noises are brought, which will affect and disturb the motion shape. The second columns in **Fig. 3** display the corresponding to **MHI**. From them, we can see that **MHI** is not so clear to describe the motion shape, which will be difficult for descriptors to represent the motions. Fortunately, the depth information is very helpful to detect the target, and we can employ depth information to detect the target and filter those static noise pixels, thus, we can achieve almost precise **MHI**.

In details, RGB image firstly is limited and filtered by depth image, and then we compute the maximum and minimum values of each pixel in video sequence, after that, the subtraction between maximum and minimum image pixels is calculated to obtain MHI for RGB channel, called **RDMHI**. The third and fourth columns in **Fig. 3** show their results respectively. The defination of the processing is shown as follows:

$$rd(i,j,t) = r(i,j,t) * d(i,j,t), t \in [1...N] \tag{1}$$

$$RDMHI(i,j) = \max\{rd(i,j,t)\} - \min\{rd(i,j,t)\}, rd(i,j,t) \neq 0, t \in [1...N] \tag{2}$$

where i and j is the pixel index, t is frame index, and N is the total frame number of an action sequence, $d(i,j,t)$ and $r(i,j,t)$ is current pixel depth value and RGB value in t frame respectively. $r(i,j,t)$ is filtered by $d(i,j,t)$ which can be calculated to obtain $rd(i,j,t)$. and $RDMHI(i,j)$ is MHI on RGB modality. **Fig. 3** demonstrates that the motion silhouette of **RDMHI** is much clearer than that of **MHI**, at the same time, a lot of static pixels are filtered in **RDMHI**.

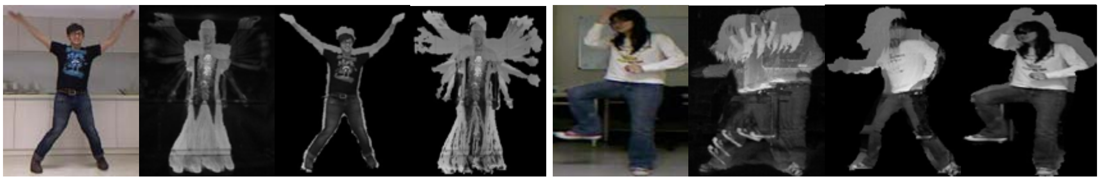


Fig.3 there are two big columns: left four and right four small columns. From left to right, there are RGB Image, traditional MHI, RGB filtered by depth and RDMHI of jacking and tai-chi actions respectively

3.2 MHI for Depth Modality

In the construction of **MHI** for RGB modality, detecting and locating the targets is difficult in real conditions, however, depth image, whose pixel values means the distance information, can be utilized to detect the targets, because the object often locate within a certain distance from the background, thus, according to the depth information, we set a suitable threshold depth value to remove the background pixels in depth image, whose depth is much bigger than the threshold, and foreground pixels in depth image are kept. Because motion history images for action representation have achieved good performance on RGB channel [1] and Megavannan *et al.* [*17*] also constructed different kind of **MHIs** and extracted different features, after that, they linked different features, which have achieved satisfying results. Thus, inspired by them, we also develop depth motion history image for all filtered depth image sequence to represent the spatial and temporal information about an action.

Suppose a depth motion image sequence $S = \{d(i,j,t)\}_1^N$, we firstly compute the maximum and minimum motion energy for each pixel in a random length N depth image sequences, and then calculate the difference between maximum and minimum images. The definitions are shown in details as follows:

$$DMHI(i,j) = \max\{d(i,j,t)\} - \min\{d(i,j,t)\}, d(i,j,t) \neq 0, t \in [1...N] \quad (3)$$

where i and j is pixel index, t is frame index, and N is the total frame number of an action sequence, $d(i,j,t)$ is current pixel depth value in t frame. In **DMHI** image, they not only convey important shape and motion clues of a human movement, but also they can filter most of static pixels. **Fig. 4** shows the corresponding RGB, depth map and depth motion history images respectively. From it, we can see that DMHI also can demonstates the motion silhouettes clearly.

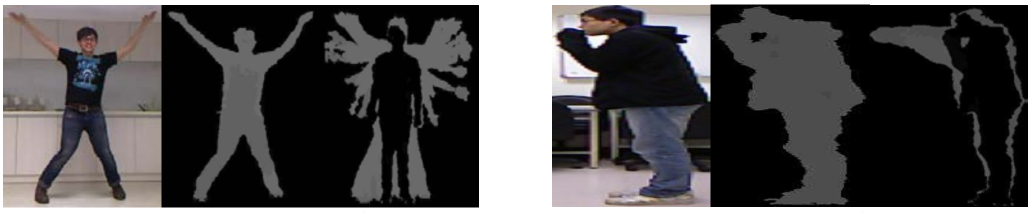


Fig. 4. there are two big columns: left three and right three small columns. From left to right: RGB Image, Depth Map and DMHI of jacking and boxing action respectively

4. Visual Representation

After we construct motion history images, the dynamic sequence of a motion became a moving human silhouette, and if we want to recognize the motion, we firstly need to describe it, thus, three different kinds of descriptors are proposed, and the following will present them in detail.

4.1 Hierarchical Blocks Descriptor

After obtaining the **DMHI** and **RDMHI**, we need to design some suitable descriptors to represent the motion. As the spatial information is very helpful for action recognition, thus, average value in hierarchical blocks (**AHB**) descriptor is proposed. In details, **DMHI** and **RDMHI** are divided into 8*8, 4*4, 2*2, 2*1 and 1*2 hierarchical blocks respectively, and then the average value with nonzero value in each block is calculated, thus, in total, the feature dimension is 88, and these descriptors are called as **DMHI_GAHB** and **RDMHI_GAHB** descriptors respectively. In addition, in order to extract accurate feature, we also firstly find a rectangular bounding box for **DMHI** and **RDMHI** images, and then employ the same scheme to split the image and extract the feature, and we name them as **DMHI_BAHB** and **RDMHI_BAHB** descriptors respectively**.** Experimental results shows the performances of **DMHI_BAHB** and **RDMHI_BAHB** are much better than that of **DMHI_GAHB** and **RDMHI_GAHB**. Thus, in the following sections, we will adopt the rectangular bounding box scheme.

4.2 Multi-Scale and Multi-Orient Descriptor

Although the spatial information is useful, the orient and scale information are also helpful, as objects often perform the same actions by different orients and different distances. Thus, we think the descriptors not only should have the spatial structure, but also should have the scale and orient information. At the same time, though a lot of perceptual experiments, researcher [42] proposed that perceptual dimensions (naturalness, openness, roughness, expansion, ruggedness) were very important for scene representation, they also employed the different scales and orients filter to compute these perceptual dimensions, in which each dimension depicts a meaningful property of the space of the scene. Large scale experiments on scene recognition showed its accuracy was very excellent. Inspired by this, we think that gist descriptor is also very helpful for our task. In fact, **DMHI** and **RDMHI** images also can be considered as different characteristics scenes in which naturalness, openness, roughness, expansion and ruggedness are very different. For example, the roughness dimension of two hands waving will be much bigger than that of one hand waving, the openness dimension of running will be larger than that of jumping. Thus, after obtaining **DMHI** and **RDMHI**, the gist descriptor is adopted, called **DMHI_GIST** and **RDMHI_GIST** descriptors respectively. In these descriptors, four scales and eight orients filter are employed, and then each filtered **DMHI** and **RDMHI** is divided into 4*4 blocks, and average value of each block is computed, thus, the dimension of **DMHI_GIST** and **RDMHI_GIST** descriptor are 512. Experiments show that these descriptors outperform some the-state-of-the-art descriptors, such as 7 Hu moment shape features, which are robust for a translation- and scale-invariant manner.

4.3 Pyramid Histogram of Orientated Gradients Descriptor

In above two kinds of descriptors, although the spatial information, multi-scale and multi-orient are employed, the average value in each block is calculated in which the neighbor information is ignored, and its description ability is limit, thus, we suppose that we should not only adopt the spatial information, but also employ much more robust descriptors which are related to the neighborhood. At the same time, histogram of orientated gradients **(HOG)** [43], in which the edge or gradient distribution of the local region is extracted, and the edge or gradient structure of the target can be represented well, was proposed to describe the human shape information, and has achieved good performance. Although gradient orients in HOG have actually given the spatial location information, the effect of different spatial partition to the performance of classification is neglected, thus, pyramid histogram of orientated gradients (**PHOG**) [44] was proposed. In fact, **PHOG**, which not only can represents the whole shape information, but also describes the local shape information and spatial relationship, is a spatial shape descriptor, and has attained satisfying performance on object classification.

Based on above analysis, we think that **PHOG** also will be very useful for our task, thus, **PHOG** is proposed to represent the human motion maps. In our task, **PHOG** not only describes the shape information of the human action, but also describes the space information of it in which both shape information and spatial information are very helpful for our tasks. **PHOG** extraction in our task can be given as follows in detail: 1) **DMHI** and **RDMHI** motion maps are constructed and calculated, and then the rectangular bounding boxes of them are searched and obtained, and background noise pixel are filtered; 2) On the basic of the rectangular bounding box, **PHOG** feature are extracted for **DMHI** and **RDMHI** motion maps respectively, and we called them as **DMHI_PHOG** and **RDMHI_PHOG** respectively. In the calculating **PHOG**, three layers pyramid image are constructed, and the range of gradient direction is 0~360 degree in each layer, and then all pixel gradient directions in each layer or each block are normalized into 20 dimensions by the weight of pixel gradients; 3) After that, the features in each layer are cascaded into **PHOG** feature, whose dimension is 1700.

5. Multi-modality Features Fusion by Collaborative Multi-task Learning

Since both RGB and depth image sequences for one action can be synchronously captured by Kinect, it is reasonably assumed that there exists intrinsically correlation among multiple modalities. Consequently, we can formulate the action recognition task with **m**ulti-**m**odality feature **f**usion as a **c**ollaborative **m**ulti-task **l**earning (**MMFCML**) problem to discover the underlying common knowledge among different modalities and consequently boost the performances. We propose to formulate the **MMFCML** problem with multimodal signal by transfer learning.

5.1. Problem Formulation

In multi-modality features fusion and **c**ollaborative multi-task learning, we are given a training set of $\bar{X} = \{X_i\}_{i=1}^{N}$ for K action classes with N training samples. Each member of $\bar{X}$ contains multi-model features $X_i = \{X_{ij}\}_{j=1}^{T} \in R^{d \times T}$ for one sample where $X_{ij} \in R^d$ means the feature representation in j^{td} modality for i^{td} and T denotes the total number of modalities. In real conditions, we often obtain limited number of samples in each action classes, thus, we collectively construct the dictionary with all classes of action samples for each modality feature. For simplicity, we define $\Phi = \{\Phi_j\}_{j=1}^{T}$ as the dictionary with multi-model bases for the **MMFCML** task, where the dictionary Φ_j for j^{td} modality, and $\Phi_j = \{X_{ij}\}_{i=1}^{T} \in R^{d \times T}$ consists of all training samples in the j^{td} modality for all action classes. For a test sample $\mathrm{Y} = \{Y_j\}_{j=1}^{T} \in R^{d \times T}$ where Y_j denotes the feature representation in the j^{td} modality, we can formulate the objective function:

$$W^* = \arg\min_{W} \sum_{j=1}^{T} \lambda_j \| Y_j - \Phi_j \times W \|_2^2 + \lambda \| W \|_2^2 \tag{4}$$

Where first term means the empirical loss function, and λ_j means the weight for the reconstructive error by the j^{th} modality feature, and this empirical loss function is formulated based on sparse coding principle. Given Y_j and the corresponding dictionary Φ_j, sparse coding means to decompose Y_j over Φ_j such that $Y_j = \Phi_j \times W + r_j$ where W is the sparse vector and r_j is the residual. W expilicitly represents the similarity between Y_j and each base of Φ_j. This term evaluates the reconstructive errors with T modality signals. The minimum of this term can lead to the latent correlation transferring among multiple modality features. At the same time, the second term consists of the rigid penalty, and λ controls the weight for regularization. If two bases are highly similar to each other, they should be assigned with almost the same weights. It is well known that the ridge penalty with strict convexity can preserve consistence for the decomposed coefficient. Therefore, the rigid penalty can impose consistence for **MMFCML**. Furthermore, it can make the least square solution of the empirical loss function stable as well as induce a certain amount of sparsity.

The advantage of the proposed **MMFCML** problem is that the consistence constraint can be incorporated with the loss function decoupled for single modality-based learning problem. This combination will facilitate transferring the common knowledge among multiple modalities to boost the performance.

5.2 Solution and Inference

Although both empirical loss function and rigid penalty are convex in W, but the L1-norm term in the objective function Eq.4 is not differentiable, thus, gradient descent method is not available for solution. However, Nesterov's method [48] utilizes a linear combination of previous two points as the search point to achieve high convergence speed. The Nesterov's method is based on two sequences $\{x_i\}$ and $\{s_i\}$ in which $\{x_i\}$ is the sequence of approximate solutions, and $\{s_i\}$ is the sequence of search points. The search point s_i is the affine combination of x_{i-1} and x_i as

$$s_i = x_i + \alpha_i (x_i - x_{i-1}) \tag{5}$$

Where α_i is the combination coefficient, and the approximate solution x_{i+1} is computed as a "gradient" step of s_i as

$$x_{i+1} = \pi_G (s_i - (1/\gamma_i) * g'(s_i)) \tag{6}$$

where $\pi_G(v)$ is the Euclidean projection of v onto the convex set G:

$$\pi_G(v) = \min_{x \in G} \frac{1}{2} \| x - v \|^2 \tag{7}$$

$1/\gamma_i$ is the step size, and γ_i is determined by the line search according to the Armijo-Goldstein rule so that γ_i should be appropriate for s_i, and the details can be found in [48]. Thus, in our paper, we adopt the Nesterov's method to solve the optimization problem in Eq.(4), and then we can derive the analytical solution as:

$$W^* = [\sum_{j=1}^{T} \lambda_j \Phi_j^T \Phi_j + \lambda I]^{-1} [\sum_{j=1}^{T} \lambda_j \Phi_j^T Y_j] \tag{8}$$

With the optimal W^*, we can compute the error by the q class as follows:

$$error(q) = \sum_{j=1}^{T} \lambda_j \| Y_j - \Phi_j^q \cdot W^{*q} \|_2^2 + \lambda \| W^{*q} \|_2^2 \tag{9}$$

Where Φ_j^q denotes corresponding dictionary for j^{td} modality feature and q action class, and the class of the test sample can be inferred by choosing the action class q with the minimum $error(q)$. For the complexity of the proposed algorithm, since $[\sum_{j=1}^{T} \lambda_j \Phi_j^T \Phi_j + \lambda I]^{-1}$ is independent of the test sample Y_j and can be pre-computed with the constructed dictionary Φ, the optimal solution of the first term in Eq.(4) can be rapidly obtained simply by projecting $[\sum_{j=1}^{T} \lambda_j \Phi_j^T Y_j]$ onto $[\sum_{j=1}^{T} \lambda_j \Phi_j^T \Phi_j + \lambda I]^{-1}$, thus, it can be computed quickly.

6. Experimental Evaluation and Discussion

In order to evaluate our proposed descriptors adequately, we will assess them by different video channels and different classification models. Although MSR-Action3D dataset [10] is a public action dataset, but it just includes depth sequences captured by a depth camera, and RGB channel is ignored, but a challenge and public DHA action dataset [9] includes both depth and RGB information together. Thus, in our experiments, DHA dataset is employed, and the popular classification models- KNN and SVM are borrowed to recognize human action, in addition, the mode-free models - SRC and CRC, also are adopted to identify human motions. In all experiments, SVM model are learned using cross-validation, and the parameters in **SRC, CRC and MMFCML** models are selected by cross validation within the range of [1, 0.5, 0.1, 0.05, 0.01, 0.005, 0.001, 0.0005, and 0.0001].

6.1 Experimental Setting Up

DHA Dataset: In this dataset [9], it contains 17 action categories: (1) bend, (2) jack, (3) jump, (4)one-hand-wave, (5) pjump, (6)run, (7)side, (8) skip, (9) two-hand-wave, (10) walk, (11)clap-front, (12) arm-swing, (13) kick-leg, (14) pitch, (15) swing, (16) boxing and (17) tai-chi, and each action was performed by 21 people (12 males and 9 females), such that there are totally 357 videos in DHA dataset, each with both the color and depth data recorded. Although the background in DHA dataset is relative clean, there are some similar actions which will be very difficult to recognize.

Features: In our experiments, we not only extract the depth features, such as **DMHI_BAHB**, **DMHI_GIST** and **DMHI_PHOG**, but also extract RGB features, such as **RDMHI_BAHB**, **RDMHI_GIST** and **RDMHI_PHOG**. In order to compare with other descriptors, translation, scale and orientation invariant 7 Hu moments, Gabor feature and LBP [46] feature are also extracted for both depth and RGB channels respectively, and these features are named as **DMHI_7_Hu_moment, DMHI_Gabor, DMHI_LBP, RDMHI_7_Hu_Moment**, **RDMHI_Gabor** and **RDMHI_LBP** respectively. In addition, we also directly fuse these descriptors to evaluate them further.

Classifiers: To assess the performance of our proposed scheme well, KNN, SVM, SRC and CRC models are constructed. For SVM with RBF kernel [47] are trained on training dataset, and then the performance is assessed on testing dataset. For SRC and CRC models, all training samples will be adopted directly as the basic vectors, and then a testing sample is reconstructed and represented by these basic vectors, after that, impulse function is designed for each class and representation, and then the minimum representation error is adopted to classify the testing sample.

Evaluation Criteria: In our previous work [48], we had discussed the evaluation protocol of action recognition, and the leave-one-person-out method should be much more reasonable. Thus, we will utilize the leave-one-person-out protocol, and the popular average accuracy is employed.

6.2 Performance Evaluation on Depth Channel

Firstly, we will evaluate our descriptors on depth channel in DHA dataset with different classification models respectively. At the same time, we also compare their performances with translation, scale and orientation invariant 7 Hu moments, Gabor and LBP descriptors. In order to compare fairly, all experimental settings are required to the same, and their performances are shown in **Table 1**.

Table 1. Performance comparison on depth channel in DHA dataset when different descriptors and models are adopted

Schemes	*RBF*	*KNN*	*SRC*	*CRC*
DMHI_PHOG	**92.4**	**79.3**	**90.6**	**89.8**
DMHI_GIST	85	76.2	82.1	0.86
DMHI_BAHB	81	76.5	81	79
DMHI_GAHB	51	46	55	54
DMHI_Gabor	65	46.5	56.9	57.4
DMHI_7_Hu_moment	44	24.3	24.4	14.9
DMHI_LBP	55.9	48.1	51	40.1

Experimental results demonstrate that no matter what kinds of models are used, the performances of **DMHI_BAHB**, **DMHI_GIST** and **DMHI_PHOG** descriptors are much better than that of **DMHI_7_Hu_Moment, DMHI_Gabor** and **DMHI_LBP** descriptors. Meanwhile, we also can observe that although **DMHI_BAHB** descriptor has the spatial information, the orient information is ignored, thus, when **DMHI_GIST** descriptor with multi-scale, multi-orient and spatial information are considered together, their performances of different models is improving. In addition, in **DMHI_BAHB** and **DMHI_GIST** descriptors, the average value in each block is adopted to represent human actions, but their neighbor information of the center pixel is ignored, thus, their performances are not so good. However, in **DMHI_PHOG** descriptor, the pixel gradient orients with the weight of pixel gradient in each block are hired to describe human motions, whose representation ability is much more robust and efficient than that of average value in each block. No matter what kinds of models are employed, **DMHI_PHOG** descriptor achieves the best performance.

In addition, we also evaluate and compare **DMHI_BAHB** and **DMHI_GAHB** descriptors in our experiments. From the comparison results, we can understand that when we split the image into blocks without rectangular bounding box, its performance just achieve about 50%, but when the rectangular bounding box is employed, its performance can improve to 80%. In other word, the rectangular bounding box is very helpful for our task, thus, in our latter experiments, we will always adopt the rectangular bounding box scheme in extraction feature.

6.3 The Assistance of Depth Information for RGB Channel

In obtaining RGB motion history maps, it often be affected by the background pixel, and the second columns in **Fig.3** has proved it. If we can acquire the foreground target, its motion history maps will be much more clear and discriminative than that of motion history maps affected by background pixels. However, in real conditions, it is difficult to segment the foreground target. Luckily, the distance between the target pixels and background pixels are very different, what is more, the depth information is just enough distance information, which can be used to segment and locate the target, and the third columns in **Fig.4** displays the corresponding results. In order to prove the assistance of depth information for RGB channel, we choose top two descriptors in depth channel and perform some comparable experiments on **DHA** dataset by training different models, and their results are provided in **Table 2**. **Table 2** shows that when we construct the motion history image on RGB channel by traditional methods, anyway models are employed, the performances of **RMHI_GIST** and **RMHI_PHOG** just about 60%**.** However, when the depth information is borrowed to segment the target, the performances of **RDMHI_GIST** and **RDMHI_PHOG** obtain big improvement. Especial for **RDMHI_PHOG** descriptor**,** when **SVM-RBF**, **SRC** and **CRC** models are adopted, their performances reach about 90%. That is to say, the depth information for RGB channel is very auxiliary, and in our latter experiments, the depth information be used to detect and locate our target.

Table 2. Assistance of depth information for RGB channel

Schemes	*RBF*	*KNN*	*SRC*	*CRC*
RMHI_GIST	67.5	51.4	66.8	63.4
RDMHI_GIST	**89.4**	**75.6**	**89.6**	**89.3**
RMHI_PHOG	61.8	49.9	70.6	71.2
RDMHI_PHOG	**92.7**	**79.6**	**91.3**	**91.9**

6.4 Performance Evaluation on RGB Channel

In order to assess our proposed descriptors further, we also evaluate them on RGB channel by different classification models, at the same time, we also compare with the-state-of-art schemes, and their performances are provided in **Table 3**. **Table 3** shows that **RDMHI_BAHB**, **RDMHI_GIST** and **RDMHI_PHOG** descriptors still are much better than that of **RDMHI_7_Hu_Moment**, **RDMHI_Gabor** and **RDMHI_LBP** descriptors regardless of what kinds of models are engaged. In addition, the performance of **RDMHI_GIST** descriptor is much better than that of **RDMHI_BAHB** descriptor, and the performance of **RDMHI_PHOG** descriptor is also much better than that of **RDMHI_BAHB, RDMHI_GIST** descriptor. That is to say, experimental results prove further that the pixel gradient orients with the weight of pixel gradient in each block are very helpful for describing human motions, whose representation ability is much more robust and efficient than that of average value in each block. What is more, a lot of descriptors are proposed [9], but the best performance is 87%. However, for our proposed **RDMHI_PHOG** descriptor, when **SVM-RBF**, **SRC** and **CRC** are employed, all their performances reach above 91%.

Table 3. Performance comparison of different Descriptors

Schemes	*RBF*	*KNN*	*SRC*	*CRC*
RDMHI_PHOG	**92.7**	**79.6**	**91.3**	**91.9**
RDMHI_GIST	89.4	75.6	89.6	89.3
RDMHI_BAHB	**77**	74.8	88.3	0.773
RDMHI_Gabor	66	31.3	48.7	67
RDMHI_7_Hu_moment	47	17.4	20	45
RDMHI_LBP	42.4	31	51	23.8

From above evaluation and analysis, we can conclude that our proposed and adopted descriptors can achieve much better performance than that of some the-state-of-the-art schemes on both depth and RGB channels, even different models are employed. What is more, the performance of RGB channel can obtain big improvement by the assisting of depth information, which is an important complement for RGB channel.

6.5 Performance Evaluation of Direct Fusion Different Modality Features

Experimental results have proved that our proposed descriptors obtain good performance on both depth and RGB channels no matter what kinds of models are adopted, but as different descriptors and different channels have some complement, thus, if we can fuse them directly, it may be helpful for action recognition. Thus, in this section, we will try to fuse our descriptors, and then compare with individual descriptor to further prove the superiority of them. In our experiments, top three descriptors are employed, and the paired descriptors among them are directly linked to form the fusion descriptor, for example, when R**DMHI_GIST** and **DMHI_BAHB** descriptors are fused, the combined descriptor is called **DMHI_BAHB_RDMHI_GIST,** and when R**DMHI_PHOG** and **DMHI_PHOG** descriptors are combined, the new fused descriptor is labeled as **DMHI_RDMHI_PHOG. A**t the same time, **KNN, SVM** with RBF and linear kernels, **SRC** and **CRC** models are adopted respectively, and their results are shown in **Fig. 5**, **Fig. 6**, **Fig. 7** and **Fig. 8** respectively. From them we can see that when descriptors in depth and RGB channels are combined, their performances are much better than that of sole descriptor, for example, when **KNN** model are used, the performances of **DMHI_PHOG** and **RDMHI_GIST** descriptors achieve 79.3% and 75.6% respectively, but the performance of **DMHI_PHOG_RDMHI_GIST** descriptor reaches 83.4%, whose improvement attains 7.8% for **RDMHI_GIST**. Similarly, when **SVM** model are trained, the fusion descriptors also can obtain a certain improvement, for example, the performances of **DMHI_PHOG, RDMHI_PHOG** and **DMHI_RDMHI_PHOG** descriptors are 92.4%, 92.2% and 96.1% respectively, and its improvement still reaches 4%. What is more, when SVM model with RBF kernel is trained, its performance keep stable and efficient, and **Fig. 6** show their results. When comparing with Lin et al. [9] (its accuracy is 87%), the improvement achieves 9.1%.

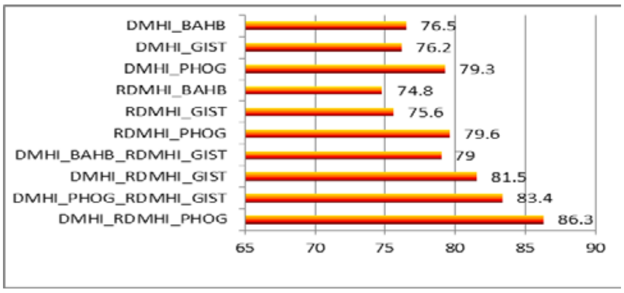


Fig. 5. Performance comparison between fusion descriptor and sole descriptor by KNN classifier

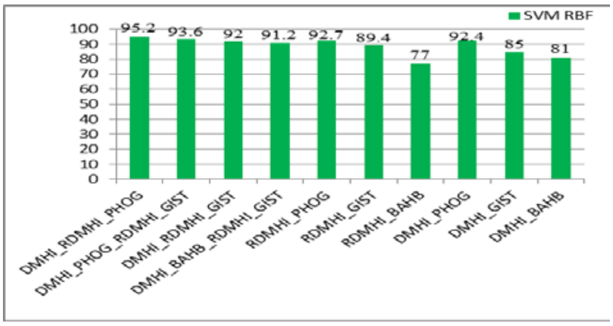


Fig. 6. Performance comparison between fusion descriptor and sole descriptor by SVM classifier

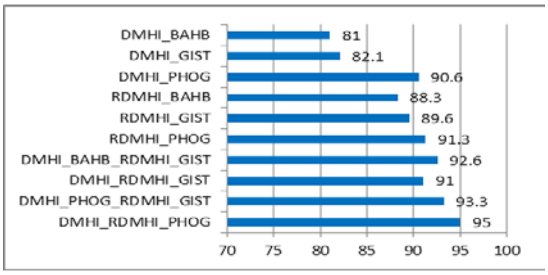


Fig. 7. Performance comparison between fusion descriptor and sole descriptor by SRC classifier

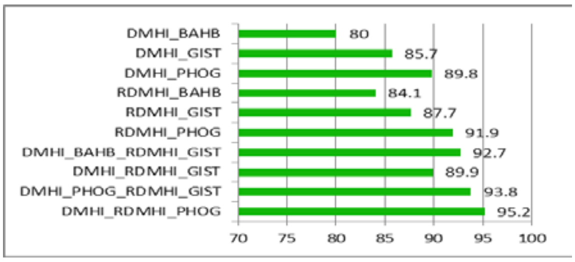


Fig. 8. Performance comparison between fusion descriptor and sole descriptor by CRC classifier

Although KNN model does not need to train the model, its best performance just reaches 86.3%, as for SVM, although its best accuracy achieves 96.1%, its training is time-consuming and its model depends on the training dataset, thus, we will employ **SRC** and **CRC** classification models, which do not depend on complicated model selection and learning, at the same time, the generalization ability of them can be easily extended by simply adding bases, the new labeled action video, to evaluate our descriptors further; **Fig. 7** and **Fig. 8** display their performances. From them, we can observe that although **SRC** and **CRC** classification models do not need to complex training, and their best accuracies achieve 95%

and 95.2 respectively, which is comparable to that of SVM model.

At the same time, we also can know that the combined descriptors are much better than that of sole descriptor no matter **SRC** and **CRC** models are adopted. For example, the accuracies of **DMHI_BAHB** and **RDMHI_GIST** descriptors are 81% and 89.6% respectively in **SRC** model, but the accuracy of fusion descriptor **DMHI_BAHB_RDMHI_GIST** is 92.6%. In addition, among all the combined descriptors, the performance of **DMHI_RDMHI_PHOG** descriptor is the best.

In conclusion, no matter what kinds of models are adopted, our combined descriptors are much better than that of sole descriptor, and their performances can obtain some improvement when comparing with sole descriptors. In addition, when comparing with the-state-of-the-art, the accuracy of our descriptor increases from 87% to 96.2%. In a word, our descriptors are robust, stable and efficient.

6.6 Performance Evaluation with the Change of Layers of Pyramid Histogram of Oriented Gadients

From above analysis, we can conclude that as neighbor gradient and pyramid scheme are applied in **DMHI_PHOG** and **RDMHI_PHOG**, thus, their accuracies are the best among all the descriptors. In order to further assess it, we will evaluate the performance variation with the change of the number of pyramid layers in **DMHI_PHOG** and **RDMHI_PHOG** descriptors by different kinds of models. Fig. 9 and Fig.10 reveal their results.

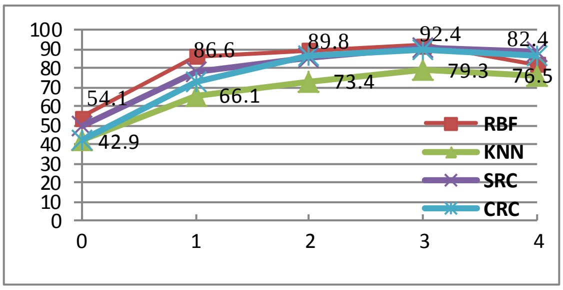


Fig. 9. Perchance evaluation on depth channel when different layers PHOG descriptors and different models are employed

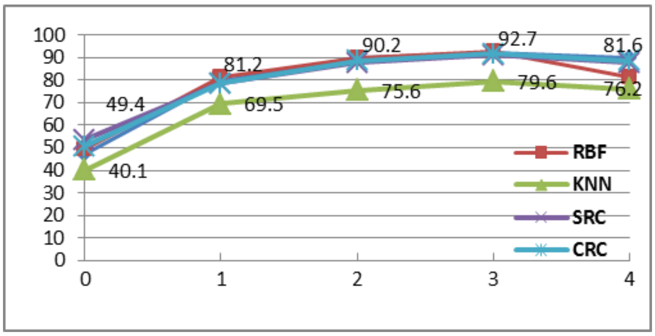


Fig. 10. Perchance evaluation on RGB channel when different layers PHOG descriptors and different models are employed

Fig. 9 demonstrates that if we just adopt oriental HOG scheme without pyramid, their accuracies range from 42.9% to 54.1%, but if the layer of pyramid is added to three, their accuracies range from 79.3% to 92.4%. That is to say, with the increase of the number of pyramid layers, their accuracies step up, but when the number of pyramid layers adds up to four, their performances drop gradually no matter what kinds of models are borrowed.

Similarly, when the evaluation is on RGB channel, the performances step up with the increase of the number of pyramid layers. And when three layers pyramid are employed, their accuracies are the best, but their performances gradual decline when the number of pyramid layers is added up.

In a word, we should adopt multi-layer in the **DMHI_PHOG** and **RDMHI_PHOG** descriptors, but the layer cannot be too big, or the accuracy will be bad. Generally speaking, three layers should be enough, whose representation ability is robust, efficient and stable.

6.7 Performance Evaluation of Collaborate Multi-task Learning

Since RGB and depth images represent one scene in different modalities, they are complementary to each other and it will benefit human action recognition by fusing both for discriminative feature representation and model construction. In section 6.5, we have proved that the concatenated different modality features is helpful, whose performances are much better than that of sole descriptor, and their performances can obtain some improvement when comparing with sole descriptors. However, the direct feature fusion of both RGB and depth information is limited to improve the performances, thus, we propose a collaborative multi-task learning method for model learning and inference based on transfer learning theory. Thus, we utilized both **DMHI_PHOG** and **RDMHI_PHOG**, **DMHI_PHOG** and **RDMHI_GIST**, **DMHI_GIST** and **RDMHI_GIST**, **DMHI_BAHB** and **RDMHI_GIST** respectively in the proposed collaborate multi-task learning framework to discover the latent correlation. To demonstrate the superiority of integrating both information of RGB and depth, we also concatenated them, and then different classifier are trained respectively and their performances are given in **Table 4**. From it, we can see that when MMFCML model is employed, we can achieve the accuracies of 97.3%, 96.1%, 95.2% and 94.9% respectively, and all of them are better than that of direct fusion scheme. When comparing with Lin et al. [9], their improvements are about 10.3%, 9.1%, 8.2% and 7.9% respectively. As for Gao et al. [50], our proposed algorithm also is comparable. The class-wise accuracy is given in **Table 5**. From it, we can see that the accuracy of most actions is above 90% and even 100%, that is to say, we almost can recognize all of actions.

Table 4. Performance evaluation and comparison of MMFCML model and others

Schemes	*KNN*	*RBF*	*SRC*	*CRC*	*MMFCML*
DMHI_RDMHI_PHOG	86.3	95.2	95	95.2	**97.3**
DMHI_PHOG_RDMHI_GIST	83.4	93.6	93.3	93.8	**96.1**
DMHI_RDMHI_GIST	81.5	92	91	89.9	**95.2**
DMHI_BAHB_RDMHI_GIST	79	91.2	92.6	92.7	**94.9**
Lin et al. [9]	87				
Gao et al. [50]	95.2				

Table 5. Class-wise accuracy when MMFCML, DMHI_PHOG and RDMHI_PHOG are employed

Action	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
Acc (%)	100	**100**	**95**	**100**	**100**	**95**	**100**	**95**	**100**
Action	10	**11**	**12**	**13**	**14**	**15**	**16**	**17**	
Acc (%)	90	**95**	**100**	**95**	**90**	**100**	**100**	**100**	

Note: 1.Bend; 2.Jack; 3. Jump; 4.One-hand-wave; 5.Pjump; 6. Run; 7.Side; 8.Skip; 9. Two-hand-wave; 10.Walk; 11.Clap-front; 12.Arm-swing; 13.Kick-leg; 14.Pitch; 15. Swing; 16.Boxing; 17.Tai-chi

7. Conclusions

In our work, human action recognition using motion maps based on pyramid histogram of oriented gradients and collaborate multi-task learning is proposed. We firstly construct the motion history image for both RGB and depth channels. At the same time, depth information is borrowed to filter RGB information. And then different action descriptors are proposed and extracted in **DMHI** and **RDMHI** to represent these actions. After that, different modality descriptors are combined by direct fusion scheme and collaborate multi-task learning to further represent human actions. Large-scale comparison experiments by different kinds of classification models on DHA datasets demonstrate that the representation ability of neighbor gradients is much robust and efficient than that of average value in each block, and the pyramid also is very helpful for our task. No matter what kinds of models are employed, the performances of **RDMHI_PHOG** and **DMHI_PHOG** descriptors are the best among all the descriptors, whose best accuracy reaches 92.7% and is much better than most of the-state-of-the-art schemes. When these descriptors are combined to represent motions, the performances of combined descriptors are much better than just using only sole descriptor no matter KNN, SVM, SRC, CRC and MMFCML models are employed. What is more, our proposed collaborate multi-task learning scheme can obtain the best performance, which is much more efficient than direct fusion scheme. In total, our proposed approach is robust, efficient and stable.

References

[1] A. Bobick and J. Davis, “The representation and recognition of action using temporal templates,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.23, no.3, pp. 257-267, 2001. Article (CrossRef Link)

[2] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, “Actions as space-time shapes,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.29, no.12, pp. 2247-2253, 2007. Article (CrossRef Link)

[3] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, “Behavior recognition via sparse spatio-temporal features,” in *Proc. of the IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pp. 65-72, 2005. http://ieeexplore.ieee.org/xpl/articleDetails.jsp?tp=&arnumber=1570899&queryText%3DBehavior+recognition+via+sparse+spatio-temporal+features

[4] C. Schuldt, L. Laptev and B. Caputo, “Recognizing human actions: a local SVM approach,” in *Proc. of the International Conference on Pattern Recognition*, ICPR, pp.32-36, 2004. http://ieeexplore.ieee.org/xpl/articleDetails.jsp?tp=&arnumber=1334462&queryText%3DRecognizing+human+actions%3A+a+local+SVM+approach

[5] I. Laptev and T. Lindeberg, “Space-time interest points,” in *Proc. of the International Conference Computer Vision*, ICCV, pp. 432-439, 2003. http://ieeexplore.ieee.org/xpl/articleDetails.jsp?tp=&arnumber=1238378&queryText%3DSpace-time+interest+points

[6] M.-Y. Chen and A.-G. Hauptmann, “MoSIFT: Reocgnizing Human Actions in Surveillance Videos,” CMU-CS-09-161, Carnegie Mellon University, 2009. http://www.cs.cmu.edu/~mychen/publication/ChenMoSIFTCMU09.pdf

[7] M. Hu, “Visual pattern recognition by moment invariants,” *IRE Transactions on Information Theory*, vol.8, no.2, pp.179-187, 1962. Article (CrossRef Link)

[8] R. Mehrotra, “Gabor filter-based edge detection,” *Pattern Recognition*, vol.25, no.12, pp. 1479-1494, 1992. Article (CrossRef Link)

[9] Y.-C. Lin, M.-C. Hua, W-.H. Cheng, Y.-H. Hsieh, H.-M. Chen, "Human Action Recognition and Retrieval Using Sole Depth Information," in *Proc. of the 20th ACM international conference on Multimedia*, pp.1053-1056, 2012.

[10] W. Li, Z. Zhang, and Z.-C. Liu, "Action recognition based on a bag of 3D points," in *Proc. of International Conference on Human Communicative Behavior Analysis Workshop*, CVPR 2010, pp.2-6. http://ieeexplore.ieee.org/xpl/articleDetails.jsp?tp=&arnumber=5543273&queryText%3DAction+recognition+based+on+a+bag+of+3D+points%2C

[11] J. W. Davis and A. Tyagi, "Minimal-latency human action recognition using reliable-inference," *Image and Vision Computing*, vol.24, no.5, pp.455–472, 2006. http://www.cse.ohio-state.edu/~jwdavis/Publications/ivc06.pdf

[12] A. A. Efros, A. C. Berg, G.Mori, and J.Malik, "Recognizing action at a distance," in *Proc. of IEEE International Conference on Computer Vision*, pp.1, 2, 2003. Article (CrossRef Link)

[13] J. L. B. D. J. Fleet and S. S. Beauchemin, "Performance of optical flow techniques," *International Journal of Computer Vision*, vol.12, no.1, pp.43-77, 1994. http://link.springer.com/article/10.1007%2FBF01420984

[14] M. J. Black, Y. Yacoob, A. D. Jepson, and D. J. Fleet, "Learning parameterized models of image motion," in *Proc. of IEEE International Conference on Computer Vision and Pattern Recognition*, pp.561-567, 1997. 1, 2. Article (CrossRef Link)

[15] A. Klaser, M. Marszalek, and C. Schmid, "A spatio-temporal descriptor based on 3d gradients," in *Proc. of The British Machine Vision Conference*, 2008. 2 http://lear.inrialpes.fr/pubs/2008/KMS08/

[16] J. Wang, Z.-C. Liu, Y. Wu, J.-S Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, CVPR, pp.1290 -1297, 2012. http://ieeexplore.ieee.org/xpl/articleDetails.jsp?tp=&arnumber=6247813&queryText%3DMining+actionlet+ensemble+for+action+recognition+with+depth+cameras

[17] V. Megavannan, B Agarwal R. Venkatesh Babu, "Human Action Recognition using Depth Maps," in *Proc. of International Conference on Signal Processing and Communications*, SPCOM pp.1-5, 2012. http://ieeexplore.ieee.org/xpl/articleDetails.jsp?tp=&arnumber=6290032&queryText%3DHuman+Action+Recognition+using+Depth+Maps

[18] Meng Wang, Hao Li, Dacheng Tao, Ke Lu, Xindong Wu, "Multimodal Graph-Based Reranking for Web Image Search," *IEEE Transactions on Image Processing*, vol. 21, no. 11, pp. 4649-4661, 2012. Article (CrossRef Link)

[19] Meng Wang and Xian-Sheng Hua, "Active Learning in Multimedia Annotation and Retrieval: A Survey," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 2, pp.10-31, 2011. http://dl.acm.org/citation.cfm?id=1899414

[20] Yue Gao, Meng Wang, Zhengjun Zha, Jialie Shen, Xuelong Li, Xindong Wu, "Visual-Textual Joint Relevance Learning for Tag-Based Social Image Search," *IEEE Transactions on Image Processing*, vol.22, no.1, pp. 363-376, 2013. Article (CrossRef Link)

[21] Meng Wang, Xian-Sheng Hua, Jinhui Tang, Richang Hong, "Beyond Distance Measurement: Constructing Neighborhood Similarity for Video Annotation," *IEEE Transactions on Multimedia*, vol. 11, no. 3, pp. 465-476, 2009. Article (CrossRef Link)

[22] Meng Wang, Xian-Sheng Hua, Richang Hong, Jinhui Tang, Guo-Jun Qi, Yan Song, "Unified Video Annotation Via Multi-Graph Learning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 19, no. 5, pp. 733-746, 2009. Article (CrossRef Link)

[23] Meng Wang, Bingbing Ni, Xian-Sheng Hua, Tat-Seng Chua, "Assistive Tagging: A Survey of Multimedia Tagging with Human-Computer Joint Exploration," A*CM Computing Surveys*, vol. 4, no. 4, Article 25, 2012. http://www.medsci.cn/sci/show_paper.asp?id=d8003193194

[24] Meng Wang, Richang Hong, Guangda Li, Zheng-Jun Zha, Shuicheng Yan, Tat-Seng Chua, "Event Driven Web Video Summarization by Tag Localization and Key-Shot Identification," *IEEE Transactions on Multimedia*, vol. 14, no. 4, pp. 975-985, 2012. Article (CrossRef Link)

[25] Yue Gao, Meng Wang, Rongrong Ji, Xindong Wu, Qionghai Dai, "3D Object Retrieval with Hausdorff Distance Learning," *IEEE Transactions on Industrial Electronics*, vol. 61, no. 4, pp. 2088-2098, 2014. Article (CrossRef Link)

[26] Yue Gao, Meng Wang, Dacheng Tao, Rongrong Ji, Qionghai Dai, "3D Object Retrieval and Recognition with Hypergraph Analysis," *IEEE Transactions on Image Processing*, vol.21, no.9, pp. 4290-4303, 2012. Article (CrossRef Link)

[27] Yue Gao, Jinhui Tang, Richang Hong, Shuicheng Yan, Qionghai Dai, Naiyao Zhang, Tat-Seng Chua, "Camera Constraint-Free View-Based 3D Object Retrieval," *IEEE Transactions on Image Processing*, vol.21, no.4, pp. 2269 -2281, 2012. http://ieeexplore.ieee.org/xpl/articleDetails.jsp?tp=&arnumber=6030936&queryText%3DCamera+Constraint-Free+View-Based+3D+Object+Retrieval

[28] Yue Gao, Meng Wang, Zhengjun Zha, Qi Tian, Qionghai Dai, Naiyao Zhang, "Less is More: Efficient 3D Object Retrieval with Query View Selection," *IEEE Transactions on Multimedia*, vol.11, no.5, pp.1007-1018, 2011. Article (CrossRef Link)

[29] Yue Gao, Rongrong Ji, Longfei Zhang, Alexander Hauptmann, "Symbiotic Tracker Ensemble Towards A Unified Tracking Framework," *IEEE Transactions on Circuits and Systems for Video Technology*, 2014.

[30] Jun Yu, Meng Wang, and Dacheng Tao, "Semi-supervised Multi-view Distance Metric Learning for Cartoon Synthesis," *IEEE Transactions on Image Processing*, Vol.21, No.11, Nov, 2012. http://ieeexplore.ieee.org/xpl/articleDetails.jsp?tp=&arnumber=6236161&queryText%3DSemi-supervised+Multi-view+Distance+Metric+Learning+for+Cartoon+Synthesis

[31] Jun Yu a, Dacheng Tao, YongRui, JunCheng, "Pairwise constraints based multi-view features fusion for scene classification," *Pattern Recognition*, Vol.46, 2013, pp.483-496. http://www.sciencedirect.com/science/article/pii/S0031320312003524

[32] Jun Yu, YongRui, and Bo Chen, "Exploiting Click Constraints and Multi-view Features for Image Reranking," *IEEE Transactions on Multimedia*, Vol.16, No.1, Jan. 2014. http://ieeexplore.ieee.org/xpl/articleDetails.jsp?tp=&arnumber=6623163&queryText%3DExploiting+Click+Constraints+and+Multi-view+Features+for+Image+Reranking

[33] Jun Yu, Dongquan Liu, Dacheng Tao , and Hock Soon Seah, 2012, On Combining Multi-view Features for Cartoon Character Retrieval and Clip Synthesis, IEEE Transactions on Systems, Man and Cybernetics-Part B: Cybernetics, Vol.42, Np.5, Oct, 2012. http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6189803

[34] Hua Wang, Feiping Nie, Heng Huang, "Multi-View Clustering and Feature Learning via Structured Sparsity," *ICML*, 2013. http://jmlr.org/proceedings/papers/v28/wang13c.pdf

[35] A. Liu, and D. Han, "Spatiotemporal Sparsity Induced Similarity Measure for Human Action Recognition," *International Journal of Digital Content Technology and its Applications*, vol.4, no.5, pp. 23-37, 2010.

[36] Zan Gao, An-An Liu, Hua Zhang, Guang-ping Xu,Yan-bing Xue, "Human action recognition based on sparse representation induced by L1/L2 regulations," *ICPR*, pp. 1868-1871, 2012. http://ieeexplore.ieee.org/xpl/articleDetails.jsp?tp=&arnumber=6460518&queryText%3DHuman+action+recognition+based+on+sparse+representation+induced+by+L1%2FL2+regulations

[37] K. Guo, P. Ishwar, and J. Konrad, "Action Recognition Using Sparse Representation on Covariance Manifolds of Optical Flow," in *Proc. of 2010 Seventh IEEE International Conference on Advanced Video and Signal Based Surveillance*, pp.188-195, 2010. http://ieeexplore.ieee.org/xpl/articleDetails.jsp?tp=&arnumber=5597145&queryText%3DAction+Recognition+Using+Sparse+Representation+on+Covariance+Manifolds+of+Optical+Flow

[38] C.-H. Liu, Y. Yang, Y. Chen, "Human action recognition using sparse representation," in *Proc. of Processing of IEEE International Conference on Intelligent Computing and Intelligent Systems*, pp.184-188, 2009. http://ieeexplore.ieee.org/xpl/articleDetails.jsp?tp=&arnumber=5357701&queryText%3DHuman+action+recognition+using+sparse+representation

[39] Z. Gao, H. Zhang, G.P. Xu, Y.B. Xue, “Human Behavior Recognition Using Structured and Discriminative Sparse Representation,” *International Journal of Digital Content Technology and its Applications*, Vol.6,No.23, 2012, PP. 416-422.

[40] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma, “Robust face recognition via sparse representation,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol.31,np.2, pp. 210-227, 2009. Article (CrossRef Link)

[41] L. Zhang, M. Yang and X. Feng, “Sparse Representation or Collaborative Representation: Which Helps Face Recognition?” in *Proc. of International Conference on Computer Vision*, ICCV 2011. http://ieeexplore.ieee.org/xpl/articleDetails.jsp?tp=&arnumber=6126277&queryText%3DSparse+Representation+or+Collaborative+Representation%3A+Which+Helps+Face+Recognition%3F

[42] A. Oliva, A. Torralba, “Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope,” *International Journal of Computer Vision*, vol.42, no.3, pp.145-175, 2001. Article (CrossRef Link)

[43] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, CVPR, pp. 886- 893, 2005. http://ieeexplore.ieee.org/xpl/articleDetails.jsp?tp=&arnumber=1467360&queryText%3DHistograms+of+oriented+gradients+for+human+detection

[44] A. Bosch, M.-X. Zisserman, “Representing Shape with a Spatial Pyramid Kernel,” in *Proc. of the 6th ACM International Conference on Image and Video Retrieval*, pp.401-408, 2007. http://dl.acm.org/citation.cfm?id=1282340

[45] B.-B Ni, G. Wang, P. Moulin, “RGBD-HuDaAct: A Color-Depth Video Database for Human Daily Activity Recognition,” in *Proc. of International Conference on Computer Vision workshop*, ICCV, pp.1147-1153, 2012. http://ieeexplore.ieee.org/xpl/articleDetails.jsp?tp=&arnumber=6130379&queryText%3DRGBD-HuDaAct%3A+A+Color-Depth+Video+Database+for+Human+Daily+Activity+Recognition

[46] S. Marcel, Y. Rodrigue, G. Heusch, “On the Recent Use of Binary Patterns for Face Authentication,” *International Journal on Image and Video Processing Special Issue on Facial image Processing*, pp.1-8, 2007. http://publications.idiap.ch/index.php/publications/show/294

[47] C.-C. Chang, C.J. Lin, 2001, LIBSVM: a library for support vector machines. 2001, http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

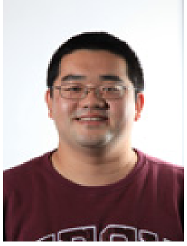
[48] Y. Nesterov, “Introductory lectures on convex optimization: A basic course,” Springer, 2004. Article (CrossRef Link)

[49] Z.Gao, M.-Y. Chen, A.-G. Hauptmann and A.-N. Cai, “Comparing Evaluation Protocols on the KTH Dataset,” in *Proc. of the First international conference on Human behavior understanding*, HBU, pp.88-100, 2010. http://link.springer.com/chapter/10.1007%2F978-3-642-14715-9_10

[50] Zan Gao, Jian-ming Song, Hua Zhang, An-An Liu, Yan-bing Xue and Guang-ping Xu, “Action Recognition Via Multi-modality Information,” *Journal of electrical engineering & Technology*, Vol.9 No. 2, pp.742-751, 2014. http://www.jeet.or.kr/LTKPSWeb/uploadfiles/be/201311/191120131352530183750.pdf

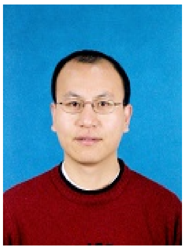
Z. Gao is an associate professor in the school of Computer and Communication engineering, Key Laboratory of Computer Vision and System, Ministry of Education, Tianjin University of Technology. From Sep. 2009 to Sep. 2010, he was a visiting scholar in the School of Computer Science, Carnegie Mellon University, USA. He received his Ph.D degree from Beijin University of Posts and Telecommunications in 2011. His research interests include computer vision, multimedia analysis and retrieval.

H. Zhang is a professor in the school of Computer and Communication Engineering, Tianjin University of Technology, Tianjin, China. She received her doctor degree from Tianjin University in 2008. Her research interests include multimedia analysis and virtual reality.



A. A. Liu is an associate professor in the school of Electronic Information Engineering, Tianjin University, P.R. China. From Sep. 2008 to Nov. 2009, he was a visiting scholar in the Robotics Institute, Carnegie Mellon University, USA. His research interests include learning-based computer vision, multimedia analysis and retrieval, biomedical image processing. He is an IEEE member

Y.B. Xue is an associate researcher in the school of Computer and Communication Engineering, Tianjin University of Technology, Tianjin, China. He received his master degree from Tianjin University of Technology in 2005. His research interests include multimedia analysis and computer vision.



G.P. Xu is an associate professor in the school of Computer and Communication Engineering, Tianjin University of Technology, Tianjin, China. He received his Ph.D and M.S degree from Nankai University in 2009 and 2005, respectively. His research interests include optimal design and performance evaluation of multimedia systems and distributed storage networks.